

Accelerating Style Transfer: Enhancing Efficiency of Diffusion-Based Models with Advanced Sampling Methods

Shenghan Zhou *

School of Big Data and Software Engineering, Chongqing University, Chongqing, China

* Corresponding Author Email: 20211979@stu.cqu.edu.cn

Abstract. Traditionally, generating images with diffusion-based style transfer models has been time-consuming. This study explores how recent advancements in accelerated sampling methods can expedite the image generation process for top-performing style transfer model. Style Injection in Diffusion (StyleID) renowned for its exceptional style transfer results, serves as the foundation for this research. The study integrates various accelerated sampling techniques—Fast Sampling of Diffusion Probabilistic Models (FastDPM), Pseudo Numerical Methods for Diffusion Models (PNDMs), DPM-Solver, and Unified Predictor-Corrector (UniPC)—into the StyleID framework. While the core techniques of StyleID remain unchanged, these accelerated methods are applied to both the inverse and reverse diffusion processes. The findings reveal that UniPC can effectively reduce the Number of Function Evaluations (NFE) from 50, as used by Denoising Diffusion Implicit Models (DDIM) in StyleID, to just 10, while maintaining the quality of the stylized images. Additionally, within the 10-50 NFE range, DPM-Solver and PNDMs yield superior results compared to DDIM. Conversely, FastDPM does not enhance the efficiency or quality of image generation at any NFE. Consequently, the use of UniPC in the sampling method results in a fivefold increase in efficiency for producing high-quality stylized images.

Keywords: Diffusion-based Style Transfer; Accelerated Sampling Methods; StyleID; Image Generation.

1. Introduction

In computer vision, style transfer is a prominent area that involves applying the artistic style of one image, like a painting style, to the original image, while maintaining the original image's content. By utilizing neural networks, it extracts the style from a style image and applies it to the original image, thereby achieving artistic effects. Style transfer has various practical applications, such as efficiently stylizing photographs into cartoon-style images in animation production [1], thereby enhancing both efficiency and visual appeal. In recent years, the field of image generation has been uniquely advanced by the development of diffusion models, with several studies beginning to utilize diffusion models for style transfer by controlling the generation process [2-5]. Style Diffusion styles images by introducing Contrastive Language-Image Pretraining (CLIP) loss and style reconstruction priors, training diffusion models for style transfer [6]. Inversion-Based Style Transfer with Diffusion Models (InST) converts style images into text-guided models that create images with specific artistic styles [3]. Other methods adapt latent distributions by style image and apply denoising techniques to generate images with that particular style [2]. However, above methods require model fine-tuning for each style, which can be time-consuming.

The study introduces a training-free method using Inject Fusion, combining features from the original and target content using h-space and U-Net skip connections, enhancing image editing in pretrained unconditional diffusion models without compromising image quality [5]. However, applying such methods in latent diffusion models like Stable Diffusion (SD) remains challenging [7]. To fully leverage large-scale pretrained diffusion models, Style Injection in Diffusion (StyleID) utilizes pretrained SD and Denoising Diffusion Implicit Models (DDIM) inversion to encode content and style images as noise via inverse diffusion processes [6]. During reverse diffusion process, it substitutes style image keys and values for corresponding features of content image in self-attention (SA), preserving query information for content. Further improvements include attention temperature scaling to maintain content structure and initialize noise by Adaptive Instance Normalization (AdaIN)

to address color inconsistencies [8]. Nevertheless, achieving style transfer for a single image to a specific style still requires significant time.

StyleID is renowned for producing outstanding style transfer results. However, creating a single stylized image involves approximately 12 seconds and necessitates 50 Number of Function Evaluations (NFE), which indicates the number of steps required for image generation. This study explores the application of accelerated sampling methods to StyleID to enhance efficiency without compromising image quality. Utilizing pre-trained SD models, this research employs non-training accelerated sampling techniques, including Fast Sampling of Diffusion Probabilistic Models (FastDPM), Pseudo Numerical Methods for Diffusion Models (PNDMs), Diffusion Probabilistic Models (DPM) Solver, and Unified Predictor-Corrector (UniPC) [9-12]. By comparing the speed and quality of these methods, the study identifies the most effective approach for optimizing StyleID's performance. Results show that UniPC successfully transfers style to the content image within 10 NFE, significantly outperforming other methods in both efficiency and effectiveness. Thus, generating a high-quality stylized image with StyleID now requires only 10 NFE or fewer, marking a substantial improvement in the style transfer process.

2. Methodology

2.1. Dataset Description and Preprocessing

In this paper, the author utilizes SD as the backbone model, using techniques in StyleID to do style transfer. These techniques enable the incorporation of the style of a given style image (I^S) into a content image (I^C), resulting in a stylized image (I^{CS}). I^{CS} Retains the original content of I^C while successfully transferring the texture patterns from I^S . The process is as follows: First, the encoder encodes I^C and I^S into z_0^C and z_0^S respectively. An inverse diffusion process is then applied, adding noise to z_0^C and z_0^S from $t=0$ until $t=T$, resulting in content noise z_T^C and style noise z_T^S . For DDIM, FastDPM, and PNDMs sampling methods, DDIM inversion is used. For DPM-Solver and UniPC, the inverse diffusion process is a reverse diffusion process from $t=0$ to $t=T$. During the inversion of I^C , the Query (Q) from the SA in U-Net is preserved, while the Key (K) and Value (V) from the SA in U-Net are preserved during the inversion of I^S . Subsequently, the AdaIN method is used to compute the initial noise z_T^{CS} from the content noise and style noise. During the reverse diffusion process, this initial noise is denoised, and the Q, K, and V preserved from the inverse diffusion process are used in place of the original Q, K, and V. Finally, the decoder decodes the predicted latent representation at $t=0$ into I^{CS} . Fig. 1 illustrates the entire process.

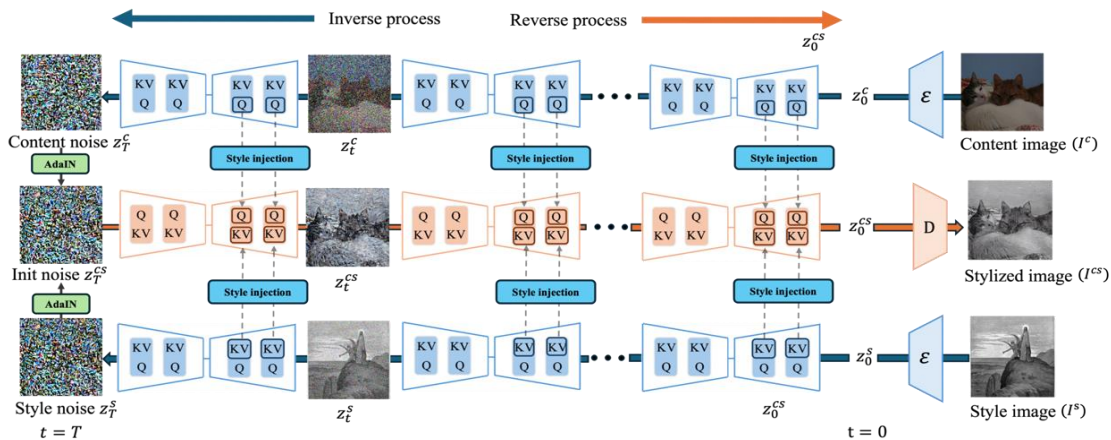


Fig 1. The pipeline of the model.

2.1.1. StyleID

SD [7] is a Latent Diffusion Model that begins by using an encoder to convert the input RGB image into a latent representation. It then employs U-Net to predict noise in the latent space.

Additionally, a decoder is used to reconstruct the image by mapping this latent representation back to the original RGB space. In addition to encoding images, SD uses a text encoder to process textual descriptions and encode it to an intermediate representation. The model utilizes a time-conditional U-Net [13] as its backbone, incorporating cross-attention and SA mechanisms. Within the U-Net's cross-attention blocks, SD applies the encoded text to calculate K and V, while the latent image representation at time t to calculate Q.

StyleID [6], based on SD, introduces three techniques to enhance style transfer performance: Attention-based Style Injection, Attention Temperature Scaling, and Initial Latent AdaIN.

Attention-based Style Injection: StyleID manipulates the attention map in SA to influence the model's output. This method can retain the content (semantics and spatial layout) of I^c while incorporating the textures of I^s . This is accomplished through the inversion diffusion process. First, in inversion diffusion process, StyleID collects the content query value (Q_t^c) from SA at each time step when I^c is input. Simultaneously, at each time step, StyleID collects the style key and style value (K_t^s, V_t^s) from SA when I^s is input. Then, StyleID utilizes the AdaIN method to compute the stylized latent noise which serves as the initial noise for the reversion process. During denoising, StyleID injects style key and style value into SA, and then the content query is combined with the query values (Q_t^{cs}) from the reverse diffusion process of the initial noise to compute the injected query value ($\tilde{Q}_t^{cs}$).

Attention Temperature Scaling: Since StyleID substitutes original K and V in SA with style keys and style values, it results in a reduced magnitude of similarity in the attention map compared to the original content and style images. Consequently, this can result in the attention map becoming blurred or smoothed, leading to the output image lacking a clear representation of both style and content information. To address this, the authors of StyleID introduce an attention temperature scaling parameter τ to control the magnitude of the attention map. The final attention weights are computed using the following formula:

$$\tilde{Q}_t^{cs} = \gamma \times Q_t^c + (1 - \gamma) \times Q_t^{cs} \quad (1)$$

$$\text{Attn}(\tilde{Q}_t^{cs}, K_t^s, V_t^s) = \text{soft max}\left(\frac{\tau \cdot \tilde{Q}_t^{cs} K_t^{sT}}{\sqrt{d}}\right) \cdot V_t^s \quad (2)$$

Where the parameter $\gamma \in [0, 1]$ is used to control the degree of content image preservation.

Initial Latent AdaIN: The Attention-based Style Injection allows for the transfer of the textures and local patterns of I^c to I^{cs} . However, the color tone of I^c is still preserved. To achieve the color tone of I^s , StyleID employs AdaIN [8] to initialize the latent noise.

$$Z_T^{cs} = \sigma(Z_T^s) \left(\frac{Z_T^c - \mu(Z_T^c)}{\sigma(Z_T^c)} \right) + \mu(Z_T^s) \quad (3)$$

Where, $\mu(\cdot)$ and $\sigma(\cdot)$ represents the channel-wise mean and variance.

2.1.2. Sampling methods

Denoising Diffusion Probabilistic Models (DDPM) is the earliest generative model that creates samples through a process of progressively adding noise and then denoising. Specifically, during training, DDPM gradually introduces noise to images and trains to reconstruct the original image from the noisy version. DDIM is an improved model based on DDPM. DDIM incorporates an implicit denoising process that decreases the number of steps required during generation, leading to improvements in both the efficiency of the process and the quality of the images.

In StyleID, the authors use DDIM for both the inversion and reversion diffusion processes. However, the sampling efficiency of DDIM remains relatively low. To address this issue, researchers have developed several fast-sampling methods that do not require additional training. These methods include FastDPM [10], PNDMs [11], DPM-Solver [12], and UniPC [9].

FastDPM introduces two sampling methods: step sampling and var sampling, aimed at improving the sampling efficiency of diffusion models. The step sampling method constructs an approximate diffusion and reverse process by selecting discrete diffusion steps. This approach does not require retraining the model; it directly leverages a pre-trained DDPM model and accelerates the process by reducing the number of sampling steps. In contrast, the var sampling method is more flexible. It constructs an approximate diffusion process by selecting a set of noise variance values. The var method utilizes the Gamma function [9] of noise levels for calculation, allowing for the selection of arbitrary noise levels and thereby constructing a shorter reverse process, which further optimizes sampling quality and efficiency. The primary difference between the two methods is that the step method relies on discrete step selection, offering less flexibility but simpler implementation, whereas the var method provides greater flexibility through the choice of noise variance values.

PNDMs establishes a theoretical connection between diffusion models and numerical methods by introducing pseudo-numerical techniques and proposes a novel approach for generating samples on specific manifolds. Specifically, PNDMs use a multi-step prediction approach to speed up the sampling process by integrating extra information at each step of the reverse diffusion process. The main concept behind this method is to apply multi-step numerical integration techniques, like pseudo Runge-Kutta methods, to strike a balance between sampling speed and the quality of the samples. This approach allows for significant reduction in the number of iterations while maintaining high-quality generation results. Empirical evidence demonstrates that PNDMs generates high-quality samples more rapidly with fewer sampling steps compared to traditional methods, such as DDIM and classical numerical methods.

DPM-Solver introduces first-order, second-order, and third-order sampling methods. The core idea is to efficiently approximate the diffusion differential equations using suitable numerical algorithms. First-Order Sampling Method (DPM-Solver-1) is based on the explicit Euler method, which performs a single linear interpolation of the differential equation. The basic step involves updating the state variable using the value at the current time step and the predicted noise. Second-Order Sampling Method (DPM-Solver-2) employs an improved midpoint method, utilizing two linear interpolations to enhance accuracy. The procedure involves introducing a midpoint between the current and the next time step, and using this midpoint's value to more precisely update the state variable. Third-Order Sampling Method (DPM-Solver-3) is based on the Heun third-order method, which builds upon the second-order approach by introducing additional midpoints to further improve sampling precision. Specifically, it uses two midpoints within each time step, applying two linear interpolations and one nonlinear interpolation to update the state variable. These three methods are essentially different order approximations of numerical methods, and the difference lies in the number and position of interpolation points. The higher order method introduces more intermediate points for interpolation, thus obtaining higher precision under phase synchronization length.

UniPC combines the Unified Predictor (UniP) and Unified Corrector (UniC) to create a framework that supports arbitrary-order accuracy. The core idea is to iteratively improve predictions by leveraging both current and previous model outputs, thereby enhancing sampling efficiency and quality. The process begins with the Unified Predictor, which estimates the next state based on current and previous states by solving the diffusion ordinary differential equations (ODEs) derived from the model noise predictions. This initial estimate is then refined by the Unified Corrector, which enhances accuracy by reusing model outputs from the current and previous time steps. The prediction and correction steps are performed iteratively, with a series of auxiliary time steps progressively refining the sample, resulting in improved sampling efficiency and precision.

In this study, the author integrates the accelerated sampling methods of FastDPM, PNDMs, DPM-Solver, and UniPC into StyleID. During the inverse diffusion process, DDIM, FastDPM, and PNDMs sampling methods employ DDIM inversion. For DPM-Solver and UniPC, the inverse diffusion process is conducted as a reverse diffusion process from $t = 0$ to $t = T$.

2.2. Evaluation Protocol

In image generation tasks, the quality of generated images is often evaluated by numerous researchers using Frechet Inception Distance (FID) as a metric. However, for style transfer tasks, FID does not account for the content of both I^c and I^s when assessing I^{cs} .

Learned Perceptual Image Patch Similarity (LPIPS) metric, introduced by [14], evaluates visual similarity between two images by computing differences in their deep neural network feature spaces.

To provide a better evaluation metric for style transfer, [15] proposes ArtFID, which combines both LPIPS and FID to assess both content preservation and style matching in generated images. Specifically, ArtFID employs the LPIPS metric to evaluate content preservation between I^{cs} and I^c , and it assesses style matching by calculating the FID between I^{cs} and I^s . A smaller ArtFID value indicates better style transfer performance. ArtFID is formally described as follows:

$$ArtFID = \frac{1 + LPIPS}{1 + FID} \quad (4)$$

3. Result and Discussion

In this section, the author will present the style transfer performance of different sampling methods under varying NFE. In Section 3.1, the datasets utilized for evaluation are presented. Section 3.2 covers the results and analysis of various sampling methods.

3.1. Dataset

Since StyleID [6] does not require training, the datasets are used solely for evaluation. In this study, the author selects 20 style images from the MS-COCO dataset and 40 content images from the WikiArt dataset. All images are center-cropped to a resolution of 512 x 512 pixels. The final model generates a total of 800 stylized images.

3.2. Results Analysis

In the results analysis, the author employs ArtFID [15] to evaluate the style transfer performance across various sampling methods. Then, the author generates stylized images using a single content image and a single style image across various 10, 12, 15, 20, and 50 NFE with different sampling methods, and visualizes the generated results.

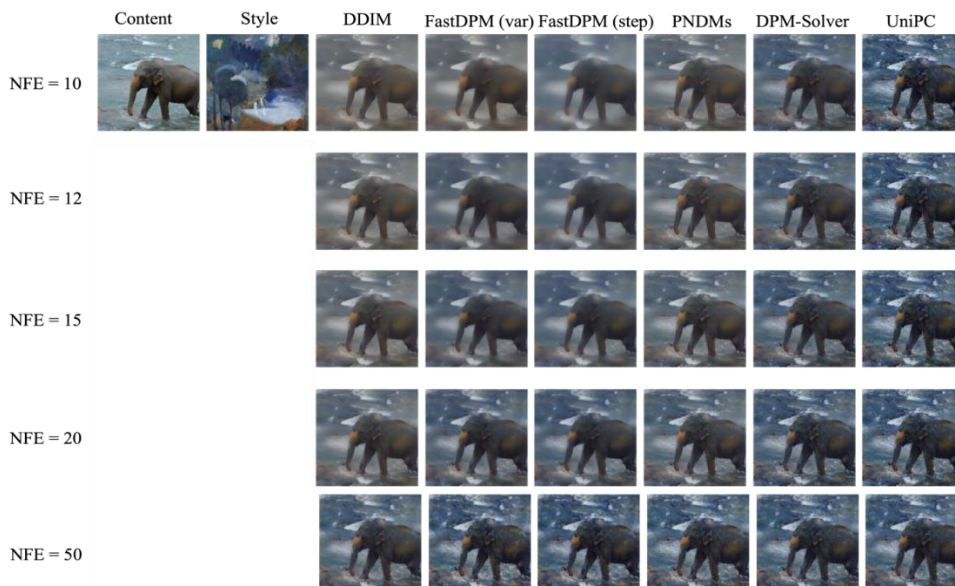


Fig 2. Samples by DDIM, FastDPM (var), FastDPM (step), PNDMs, DPM-Solver and UniPC with 10, 12, 15, 20 and 50 NFE.

Fig. 2 displays the results of generating stylized images from one content image and one style image at different NFE, with the generated images aligning with the previously discussed analysis. Fig. 3 shows the ArtFID values for different sampling methods under various NFE. The curves in the chart represent the performance of each sampling method under different NFE conditions. Lower ArtFID values indicate higher quality of the generated stylized images. The results show that as NFE increases, the ArtFID values for all sampling methods decrease and stabilize. From NFE 50 onwards, the ArtFID values of all models are nearly identical. FastDPM's step sampling and var sampling methods perform similarly, with minimal differences in ArtFID values and similar trends, starting with high ArtFID values at low NFE, indicating poorer style performance. Between 10-50 NFE, DPM-Solver outperforms PNDMs, which in turn outperforms DDIM. UniPC demonstrates the best performance among all sampling methods, converging at very low ArtFID values by 10 NFE, with little variation thereafter. This indicates that UniPC can generate images with excellent quality using just 10 NFE or even fewer, significantly enhancing the efficiency of image generation. Fig. 4 illustrates the average time required to generate a stylized image using different sampling methods at 10, 12, 15, 20, and 50 NFE. When the NFE is the same, DDIM has the shortest image generation time, while PNDMs takes the longest time at 10, 12, 15, and 20 NFE, and UniPC takes the longest time at 50 NFE. Despite the same NFE, the time differences between different sampling methods for generating an image are minimal. Since UniPC can produce a stylized image at 10 NFE with the same quality as DDIM at 50 NFE, using UniPC as the style transfer sampling method means that an excellent stylized image can be generated in approximately two seconds.

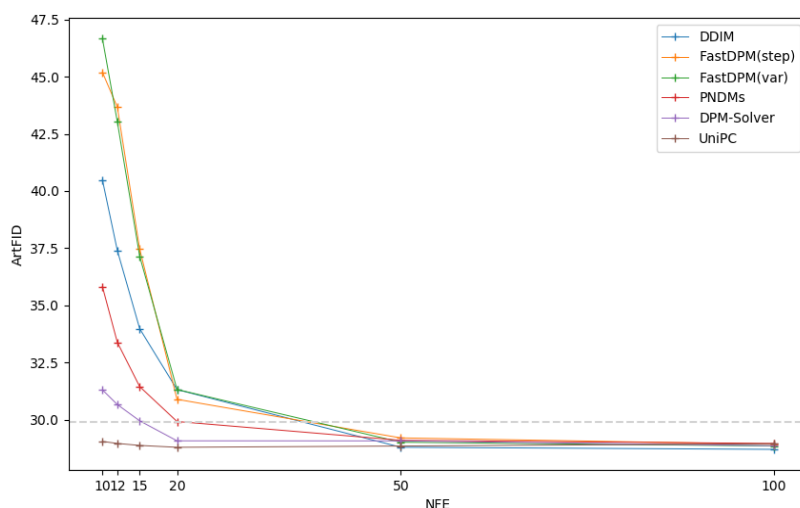


Fig 3. ArtFID with 10, 12, 15, 20, 50 and 100 number of NFE.

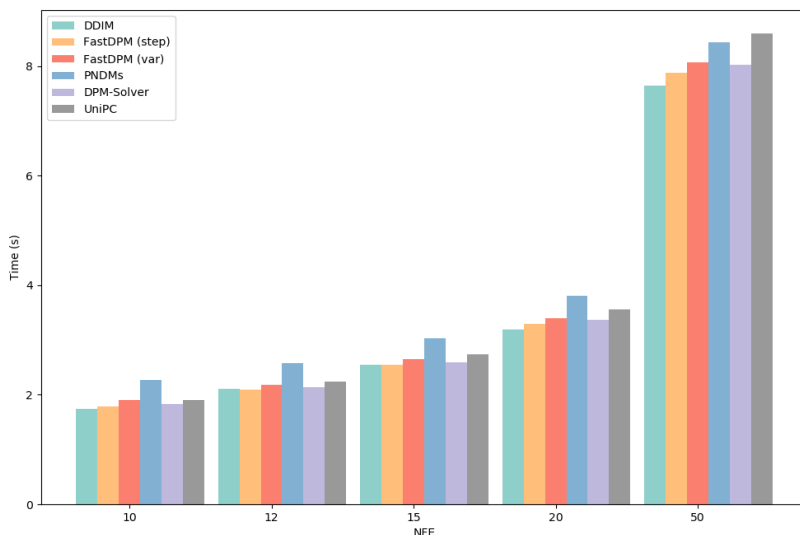


Fig 4. The average time at 10, 12, 15, 20, and 50 NFE.

In summary, the performance of StyleID is not improved by FastDPM. Instead, it diminishes the quality of stylized images and requires more NFE, significantly raising the cost of sampling an image. Compared to DDIM, DPM-Solver and PNDMs can reduce ArtFID values at the same NFE but cannot generate excellent stylized images with fewer NFE. UniPC performs the best, effectively replacing the 50 NFE originally used in StyleID with just 10 NFE or in approximately 2 seconds while achieving the same image generation quality.

4. Conclusion

Diffusion-based style transfer models traditionally require significant time to generate images. This study aims to leverage current accelerated sampling methods to enhance the image generation speed of the leading style transfer model, StyleID, known for its superior results. By integrating various accelerated sampling techniques—FastDPM, PNDMs, DPM-Solver, and UniPC—into the StyleID framework, this research seeks to optimize the image generation process. The study maintains the three core techniques proposed by StyleID while incorporating the accelerated sampling methods into both the inverse and reverse diffusion processes. Experimental results demonstrate that UniPC can replace the original 50 NFE used by DDIM in StyleID with just 10 NFE, maintaining the quality of the generated stylized images. Additionally, within the 10-50 NFE range, DPM-Solver and PNDMs outperform DDIM in generating high-quality stylized images. However, FastDPM does not enhance the model's efficiency or output quality at any NFE. Ultimately, all sampling methods converge to similar performance levels when the NFE is around 50, but UniPC significantly improves efficiency by reducing the necessary NFE to achieve comparable results.

References

- [1] Saribal C. et al. Detecting handedness of spatially oriented molecules by Coulomb explosion imaging. *The Journal of Chemical Physics*, 2021, 154(7).
- [2] Everaert. Martin N. et al. Diffusion in style. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023: 2251-2261.
- [3] Zhang Y.X. et al. Inversion-based style transfer with diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023: 10146-10156.
- [4] Wang Z.Z. Lei Z. and Wei X. Stylediffusion: Controllable disentangled style transfer via diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023: 7677-7689.
- [5] Jeong J. Mingi K. and Youngjung U. Training-free Content Injection using h-space in Diffusion Models. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024: 5151-5161.
- [6] Chung J. Sangeek H. and Jae-Pil H. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024: 8795-8805.
- [7] Rombach R. et al. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 10684-10695.
- [8] Huang X. and Serge B. Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE international conference on computer vision*. 2017: 1501-1510.
- [9] Zhao W.L. et al. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 2024, 36.
- [10] Kong, Z.F. and Wei P. On fast sampling of diffusion probabilistic models. 2021, arXiv preprint: 2106.00132.
- [11] Liu L.P. et al. Pseudo numerical methods for diffusion models on manifolds. 2022, arXiv preprint: 2202.09778.
- [12] Lu C. et al. DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 2022, 35: 5775-5787.

- [13] Ronneberger O. Philipp F. and Thomas B. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI, 2015: 234-241.
- [14] Zhang R. et al. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, 586-595.
- [15] Wright M. and Björn O. Artfid: Quantitative evaluation of neural style transfer. DAGM German Conference on Pattern Recognition. Cham: Springer International Publishing, 2022: 560-576.